

SIS | 2022

51st Scientific Meeting
of the Italian Statistical Society

Caserta, 22-24 June

Università degli Studi della Campania *Luigi Vanvitelli*

Società Italiana di Statistica

www.unicampania.it

Book of the Short Papers

Editors: Antonio Balzanella, Matilde Bini, Carlo Cavicchia, Rosanna Verde

Città di Caserta

1222-2022 800 ANNI

UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE STATISTICHE

STATA

sas

UNIVERSITÀ DEGLI STUDI DEL SANNIO Benevento

Pearson

Matilde Bini (Chair of the Program Committee) - *Università Europea di Roma*
Rosanna Verde (Chair of the Local Organizing Committee) - *Università della Campania "Luigi Vanvitelli"*

PROGRAM COMMITTEE

Matilde Bini (Chair), Giovanna Boccuzzo, Antonio Canale, Maurizio Carpita, Carlo Cavicchia, Claudio Conversano, Fabio Crescenzi, Domenico De Stefano, Lara Fontanella, Ornella Giambalvo, Gabriella Grassia - Università degli Studi di Napoli Federico II, Tiziana Laureti, Caterina Liberati, Lucio Masserini, Cira Perna, Pier Francesco Perri, Elena Pirani, Gennaro Punzo, Emanuele Raffinetti, Matteo Ruggiero, Salvatore Strozza, Rosanna Verde, Donatella Vicari.

LOCAL ORGANIZING COMMITTEE

Rosanna Verde (Chair), Antonio Balzanella, Ida Camminatiello, Lelio Campanile, Stefania Capecchi, Andrea Diana, Michele Gallo, Giuseppe Giordano, Ferdinando Grillo, Mauro Iacono, Antonio Irpino, Rosaria Lombardo, Michele Mastroianni, Fabrizio Maturo, Fiammetta Marulli, Paolo Mazzocchi, Marco Menale, Giuseppe Pandolfi, Antonella Rocca, Elvira Romano, Biagio Simonetti.

ORGANIZERS OF SPECIALIZED, SOLICITED, AND GUEST SESSIONS

Arianna Agosto, Raffaele Argiento, Massimo Aria, Rossella Berni, Rosalia Castellano, Marta Catalano, Paola Cerchiello, Francesco Maria Chelli, Enrico Ciavolino, Pier Luigi Conti, Lisa Crosato, Marusca De Castris, Giovanni De Luca, Enrico Di Bella, Daniele Durante, Maria Rosaria Ferrante, Francesca Fortuna, Giuseppe Gabrielli, Stefania Galimberti, Francesca Giambona, Francesca Greselin, Elena Grimaccia, Raffaele Guetto, Rosalba Ignaccolo, Giovanna Jona Lasinio, Eugenio Lippiello, Rosaria Lombardo, Marica Manisera, Daniela Marella, Michelangelo Misuraca, Alessia Naccarato, Alessio Pollice, Giancarlo Ragozini, Giuseppe Luca Romagnoli, Alessandra Righi, Cecilia Tomassini, Arjuna Tuzzi, Simone Vantini, Agnese Vitali, Giorgia Zaccaria.

ADDITIONAL COLLABORATORS TO THE REVIEWING ACTIVITIES

Ilaria Lucrezia Amerise, Ilaria Benedetti, Andrea Bucci, Annalisa Busetta, Francesca Condino, Anthony Cossari, Paolo Carmelo Cozzucoli, Simone Di Zio, Paolo Giudici, Antonio Irpino, Fabrizio Maturo, Elvira Romano, Annalina Sarra, Alessandro Spelta, Manuela Stranges, Pasquale Valentini, Giorgia Zaccaria.

Copyright © 2022

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891932310

Administrative database and official statistics: an IT and statistical procedure

Database amministrativi e statistiche ufficiale: una procedura statistico-informatica

Caterina Marini, Vittorio Nicolardi

Abstract The availability of a complete harmonized database that includes all information, both administrative and statistical, on the phenomenon under investigation is a precious analytical instrument very often unavailable because of issues related to important obstacles that involve both statistical and computer problems, and administrative and legal impediments. In this work, we depict the IT and Statistical procedure used to provide the first Full Information Harmonised Real Estate Database in the analysis of the real estate economy.

Abstract *La disponibilità di un database completo e armonizzato che includa tutte le informazioni, sia amministrative che statistiche, riguardanti il fenomeno in analisi è uno strumento analitico prezioso molto spesso non disponibile a causa di problemi legati a importanti ostacoli che coinvolgono sia aspetti statistici e informatici, sia impedimenti amministrativi e legali. In questo lavoro, noi descriviamo la procedura informatica e statistica usata per produrre il primo Database Immobiliare Armonizzato a Informazione Completa da usare nell'analisi del fenomeno economico del settore immobiliare.*

Key words: Administrative Data, Big Data, Real Estate Economy, Harmonised Database.

1 Introduction

[1] Caterina Marini, Department of Economics and Finance University of Bari Aldo Moro, email: caterina.marini@uniba.it

Vittorio Nicolardi, Department of Economics and Finance University of Bari Aldo Moro, email: vittorio.nicolardi@uniba.it

The debate about the administrative databases and their use to provide a statistical information that is not differently affordable is internationally considered significant and relevant to face the nowadays scientific challenge. Many issues are involved, and all are basically fundamental to provide a complete depiction of the socio-economic scenario in a world that is becoming more complex to be analysed. It is believed that the administrative data are largely available and readily usable, and this is one the most common errors that is unconsciousness committed. When working with administrative data, the main problem is the integrated use of data that are yielded by independent administrative offices, sometimes belonging to the same organisation, public authority or company, and differently collected for specific administrative purpose. Furthermore, in most cases databases are structurally diverse because of well-established IT and administrative procedures that are not updated to align all information for users. As a consequence of their nature, administrative data can be affected by material and human errors, such as duplication, missing values and erroneous information, and therefore they are not controlled and statistically correct. Another aspect that is important to consider is that administrative data are a little part of the more complex phenomenon known as Big Data. In literature, the Big Data aspect of the problem is widely discussed (Harford, (2014); Kitchin, (2014 a, b); Marini and Nicolardi, (2020) (2021)). Finally, it is crucial the harmonisation of the administrative data, which are raw information, with the official statistical data as yielded by the National Statistical Institutes (NSIs, hereafter) and, therefore, consolidated statistically. In some of our recent works, we provided an original fully comprehensive statistical tool to describe the phenomenon of the real estate economy: the full information harmonised database is referred to the real estate phenomenon where administrative data and Italian NSI's data are perfectly aligned and the Big Data crucial aspects solved. In view of the current Italian government intention to reform the Italian National Registry of private and public properties (known as "catasto") as stated in the National Recovery and Resilience Plan document as part of the Next Generation EU programme, we consider our Full Information Harmonised Real Estate Database (FIHRE-DB, hereafter) and the economic experiment we run on a sole city to be an important starting point towards a fair reform. In this work, we describe the IT and statistical procedure we defined to create the FIHRE-DB. The full procedure needed SAS software and QGIS software to yield geo-maps and overlapping steps.

2 The Data

In our work, the experiment is restricted to the territory of the city of Bari, in southern Italy, because that was part of a national research project. The databases used to create the FIHRE-DB are essentially 5 key databases (DBs, hereafter) and 3 instrumental DBs, the latter useful for harmonising all data. The 5 key DBs are independent administrative DBs, generally managed by independent Public Administration (PA, hereafter) offices of the city of Bari, and provide autonomous real estate information. Four of 5 key DBs belong to the Real Estate Registry (RER,

hereafter) and contain all information related to the real estates. The last key DB belongs to the Real Estate Italian Observatory (REIO, hereafter) of the Italian Revenue Agency. The 3 instrumental DBs are mutually fully independent because they belong to independent PA: the first belongs to RER and allows the final connection between the administrative data and the official statistical data; the second belongs to REIO and the last belongs to the Italian NSI, and both are the geo-localisation datasets used for the final overlapping action.

3 The Method

The method we proposed is a *step by step* procedure, depicted in the subsections that follow, to obtain the final outcome, that is the FIHRE-DB. A first study of all databases highlights that the 5 RER DBs needed an inevitable prior action of cleaning because of the many duplicated, missing and erroneous data. Afterwards, that was possible to proceed with the harmonisation of all data based on key fields.

3.1 RER Databases

The administrative RER DBs are fundamental to create the final harmonised database and, at the same time, those that required more effort in terms of cleaning. Furthermore, the analysis of data highlighted also that a recoding action was necessary to correct some information that, in the past, was erroneously recorded because of some administrative territorial process that involved the inclusion of some little municipalities in the city. Therefore, the cleaning and recoding actions have been important to homogenise information provided by RER and align the 5 RER DBs. The first RER DB includes a list of records that contains all technical and economic information of each unit. Therefore, without affecting the statistical significance of the analysis we deleted the incomplete and duplicated records in this first RER DB. That is a double step procedure. The first step allowed to sort data by means of two key fields: *Id_Property*, which identifies each real estate in RER, and *Ordinal_Number*, which identifies the mutations corresponding to the units:

```
proc sort data=name_Database1;
by Id_Property DESCENDING Ordinal_Number;                    [1]
```

The second step of the procedure permitted to clean the DB and utilises as key field *Id_Property* to delete the oldest data related to each real estate mutation:

```
proc sort data= name_Database1 out= name_Database2 nodupkey ;
by Id_Property;                                              [2]
```

Furthermore, we proceeded to also delete all real estate units without income:

```
data mainDatabase1;
set name_Database2;                                          [3]
where (Income not in (.));
```

The second RER DB comprises other real estate information, mainly the Urban

Section and the Registry Sheet, Subordinate and Parcel. This second RER DB includes a much higher number of records than that of the first RER DB because of both duplications, caused by some administrative change, and the presence of some real estate unit whose record has not been deleted though the building was demolished. Therefore, we deleted the incomplete and duplicated records in this second RER DB and proceeded to recoding some field. That is a multiple step procedure. The first step allowed to sort data by means of two key fields: i.e. *Id_Property* and *Urban Section*:

```
proc sort data= name_Database3;
by Id_Property DESCENDING Urban Section;                    [4]
```

The second step of the procedure permitted to clean the DB and utilises as key field *Id_Property* to delete duplicated data for each real estate:

```
proc sort data= name_Database3 out= name_Database4  nodupkey ;
by Id_Property;                                             [5]
```

Furthermore, we proceeded to also delete all records where the real estate units have an erroneous code for the field *Registry Sheet*.

```
data name_Database5;
set name_Database4;                                         [6]
where (Registry Sheet in (report the right "Registry Sheet
numbers/code" set));
```

The third step of the procedure allowed to recode the sole field *Urban Section* in this final DB:

```
data name_Database6;
set name_Database5;                                         [7]
if (Urban Section in (report the old "Section code" set)) then
Section =  report the new "Section code";
…;
```

Finally, to homogenise information between the two achieved DBs the last step deleted all the records that were in the second DB and not in the first DB. That is a *right join* between the *name_Database6* and *mainDatabase1*. This final step yielded the *mainDatabase2*. The third RER DB comprises the toponyms of each real estate unit that are important to identify the exact localisation of each unit on the urban territory. This third RER DB shows many duplications because of the modifications of some toponym and/or building number. Therefore, a cleaning action is necessary and a double step procedure was defined. The first step allowed to sort data by means of two key fields: *Id_Property*, which identifies each real estate in RER, and *Ordinal_Number*, which identifies all the mutations in toponyms and/or building number corresponding to the real estate:

```
proc sort data= name_Database7;
by Id_Property DESCENDING Ordinal_Number;                   [8]
```

The second step of the procedure permitted to clean the database and utilises as key field *Id_Property* to delete the oldest data related to each real estate mutation:

```
proc sort data= name_Database7 out= name_Database8 nodupkey ;
by Id_Property;                                             [9]
```

Finally, to homogenise information between the two previous achieved databases a final step of the procedure deleted all the real estate records that are in this third database and not in the previous two. That is a *right join* between the

name_Database8 and *mainDatabase1*. This final step yielded the *mainDatabase3*.

The final step of the procedure that is referred to the first 3 key RER DBs is the creation of a sole omni comprehensive DB that aligned data by means of the key field *Id_Property*:

```
data mainDatabaseRER;
merge mainDatabase3 mainDatabase2 mainDatabase1;
by Id_Property;
```

The last RER DB contains all data referred to the Real Estate Income and Value, both administrative data, and the corresponding council tax rate. The cleaning and recoding problems we found in this last RER DB are the same we encountered in the second RER DB previously described. In this last RER DB the codification of the real estate is based on the taxpayer in a framework of Urban Section, Registry Sheet, Subordinate and Parcel. Duplication rises because more than one taxpayer can pay the council tax for the same building in a co-ownership status. Therefore, we followed exactly the same steps of the procedure as in [5] - [7] and worked out the needed databases: first the *name_DatabaseT1* starting from the raw database; second the *name_DatabaseT2* as the outcome of the procedure [6] and [7], and finally the *mainDatabaseCT* as the result of the procedure [5].

3.2 REIO Database

The REIO DB is the last key DB we use in our work and one of the most reliable sources of data to analyse the real estate monetary value dynamics in the Italian real estate market. The REIO real estate value data are calculated based on the trade price per square meter of the properties. They are open data on a biannual basis referred to the minimum and maximum price for all the different types of real estates at the level of the council territory. To use a univocal REIO value in the analysis, the midrange value for each record was calculated. In the REIO dataset, the council territory is split into homogeneous areas that experience the same economic and socioenvironmental characteristics, i.e. the REIO zones. The REIO DB comprises statistical valued data. Therefore, no cleaning and recoding actions were necessary.

3.3 Full Information Harmonised Real Estate Database

Once the RER DBs have been achieved in a homogenous format, the last section of the work is the creation of the FIHRE-DB. The most crucial step in this last part of the procedure is to find out a method to connect and merge all the RER data with the REIO data. In fact, both *mainDatabaseRER* and *mainDatabaseCT* cannot be connected with REIO DB by any field. Therefore, the instrumental databases as describe previously play an important role to obtain the FIHRE-DB. They are all geo-referred and, therefore, a geo-localised process is suitable to solve the problem. In fact, we defined a multiple procedure of merging all the key databases through

the geo-localised information. First, the creation of a "linking table" named BRIDGEDB allowed to connect the REIO zones with the Italian NSI census sections. In other words, the procedure allows to link the Italian NSI census sections, and indirectly all real estate units, with each REIO zone and, consequentially, the statistical information with all the administrative data. Second, the multiple procedure of merging all the key databases to yield a *Final Administrative DB* was composed of the following steps:

1. merging Italian NSI census sections and *mainDatabaseRER* through the key fields *Urban Section*, *Registry Sheet* and *Subordinate*
2. merging Italian NSI census sections and *mainDatabaseCT* through the key fields *Urban Section*, *Registry Sheet* and *Subordinate*
3. merging *mainDatabaseRER* and *mainDatabaseCT* through the key field *Id_Property*

Finally, the BRIDGEDB allowed to connect the *Final Administrative DB* to the REIO DB by means of Italian NSI census sections. This is the final outcome of our work, that is FIHRE-DB.

4 Concluding remarks

The importance of the analysis yielded in this work is unique and original in its attempt to provide a new omni comprehensive database that includes all information, both administrative and statistical, related to the analysis of the real estate economy.

The experiment we run to solve the dearth of a complete information because of the practical non-existence of a full information harmonized database can be an important input to replicate the procedure for any dimensional geographical area and for any other economic analysis. As seen in this work, the IT and statistical procedure we describe to create the Full Information Harmonised Real Estate Database is quite complex because of the many issues that need to be faced and the dimension of databases. However, in view of the current Italian government intention to reform the Italian National Registry of private and public properties, our outcome can be an important starting point towards a fair reform.

References

1. Hadford, T.: Big data: a big mistake? Significance, 11, 14–19 (2014).
2. Kitchin, R.: Data, new epistemologies and paradigm shift. Big Data & Society., 1, (2014a).
3. Kitchin, R.: The Data Revolution: Big Data, Open Data. Data Infrastructures and Their Consequences: Sage Publications, London (2014b).
4. Marini C., Nicolardi V.: Big Data and Economic Analysis: The Challenge of a Harmonized Database, Data Science and Social Research II. DSSR 2019. Studies in Classification, Data Analysis, and Knowledge Organization, Springer Nature, (2020).
5. Marini C., Nicolardi V.: Administrative database and official statistics: The case of the real estate analysis. Statistica Applicata - Italian Journal of Applied Statistics - vol. 33 (1), pp.83-95 (2021).